Dean&Francis

ISSN 2959-6157

Personalized Federated Learning for Heart Failure Mortality Prediction under Non-IID Clinical Data

Zhenxuan Wang

Department of ECE, University of Illinois Urbana-Champaign, Illinois, United States
zw71@illinois.edu

Abstract:

This research explores how federated learning (FL) can be applied to predict heart failure mortality, emphasizing the protection of patient data privacy while maintaining model accuracy in realistic clinical environments. Using the Heart Failure Clinical Records dataset, I simulate 30 medical institutions and compare three FL algorithms—FedAvg, FedProx, and pFedMe—under conditions where data is not independently and identically distributed (non-IID). FedAvg serves as a baseline for centralized aggregation, FedProx introduces a proximal term to address client drift, and pFedMe employs Moreau envelopes for personalized model updates. The results demonstrate that both FedProx and pFedMe outperform FedAvg, achieving a final test accuracy of 70% versus 66.67%, with pFedMe further exhibiting the most stable training dynamics and minimal loss drift. These findings underscore the critical role of regularization and personalization in federated healthcare systems, particularly in heterogeneous environments where data distributions vary across clients. The work provides practical guidance for FL deployment in real-world medical settings, highlighting a trade-off between accuracy, model stability, and patient-specific adaptation—all while maintaining strict data privacy compliance. This contributes to the broader effort of enabling secure, collaborative AI in healthcare.

Keywords: Federated Learning; Medical AI; FedAvg; FedProx; pFedMe.

1. Introduction

In recent years, the increasing concern over the privacy of personal medical records has driven the development of machine learning techniques that respect data confidentiality while still delivering accurate and actionable predictions. Traditional centralized machine learning approaches require collecting sensitive health data—including physiological signals, body metrics, and behavioral patterns—on central servers,

posing serious privacy, security, and regulatory compliance risks. Such risks are particularly critical in healthcare settings, where patient data is governed by strict privacy laws and ethical standards, including strict regulations like HIPAA and GDPR.

Federated learning gained attention as a potential solution to privacy-related issues. By supporting collaborative training across distributed clients while not sharing raw data, FL ensures that personal information such as weight, height, age, and biometric signals remains securely stored on local devices or within institutional boundaries. This approach aligns well with privacy-preserving principles and has the potential to unlock the value of distributed medical data while minimizing exposure to privacy breaches.

However, deploying FL in healthcare contexts presents unique challenges beyond privacy. Medical data is often non-independent and identically distributed (non-IID) across clients, reflecting differences in demographics, behaviors, sensors, or institutional practices. Such heterogeneity can degrade the performance and stability of federated models. While the standard Federated Averaging (FedAvg) algorithm provides a baseline for distributed learning, it is known to perform poorly under highly non-IID conditions. To address this, several algorithmic variants have been proposed, including Federated Proximal (FedProx), which introduces a proximal term to mitigate client drift, and Personalized Federated Learning via Moreau Envelopes (pFedMe), which tailors models to each client while maintaining collaborative benefits.

This research designs and evaluates a privacy-preserving federated learning framework for predicting mortality from Heart Failure Clinical Records, with a comprehensive comparison of three aggregation algorithms: FedAvg, FedProx, and pFedMe.

The Heart Failure Clinical Records is exceptionally well-suited for federated learning (FL) research due to its clinically significant context and inherent heterogeneity. Heart failure affects over 64 million people globally and carries a 5-year mortality rate of 45-50%, making predictive modeling a high-impact application. The dataset captures 12 clinically relevant features—including demographics (age, sex), comorbidities (diabetes, hypertension), and critical biomarkers (ejection fraction, serum creatinine)—that naturally vary across healthcare institutions. These variations mirror real-world data distribution challenges in medicine, where patient demographics, diagnostic protocols, and comorbidity profiles differ substantially between hospitals (e.g., serum creatinine levels span 0.5-9.4 mg/dL across patients). This level of heterogeneity enables thorough evaluation of how well FL algorithms manage non-IID data imbalances, which remains a fundamental obstacle in medical collaborations across institutions.

The dataset's structure offers practical advantages for FL experimentation. With 299 patient records, it enables simulation of 10-50 realistic client nodes without excessive fragmentation—preserving statistical power while accommodating resource constraints. Its moderate dimensionality avoids computational bottlenecks yet retains clinical complexity, as demonstrated by centralized models achieving ~0.85 AUC. Crucially, it supports multiple non-IID partitioning strategies: clustering by comorbidity profiles (e.g., diabetes/hypertension co-occurrence), age cohorts (40-55 vs. >75 years), or biomarker ranges. This flexibility allows controlled experiments on personalization effectiveness—essential for tailoring predictions to subgroups like elderly patients with renal impairment. Furthermore, the mortality prediction task aligns with FL's privacy-preserving paradigm, as sensitive data never leaves originating "hospitals" during training.

2. Related work

Federated learning (FL) offers an effective approach for developing ML models across distributed data sources while safeguarding data privacy. McMahan et al. proposed the FedAvg algorithm to perform distributed optimization with low communication costs [1]. However, FedAvg suffers from performance degradation in the presence of non-IID data, a common characteristic of real-world medical settings [2].

To address this, FedProx was proposed by Li et al. as a modification of FedAvg, introducing a proximal term to the client objective to mitigate client drift and stabilize convergence [3]. Building further on the need for robustness in heterogeneous settings, Dinh et al. proposed pFedMe, which frames FL as a bi-level optimization problem, enabling clients to learn personalized models regularized toward a shared global model [4]. This personalization is particularly crucial in healthcare, where patient populations vary widely between institutions.

Several studies have demonstrated the effectiveness of FL in medical applications. Sheller et al. applied FL to brain tumor segmentation, showing comparable results to centralized methods while maintaining patient privacy [5]. Similarly, Li et al. used FL for COVID-19 diagnosis across hospitals using chest CT scans [6]. Moreover, Xu et al. applied FL to wearable sensor data for cardiovascular disease prediction, demonstrating the utility of decentralized learning for longitudinal health monitoring [7].

Despite these successes, many early FL applications in healthcare focused on imaging or relatively homogeneous data sources. Recent work has emphasized the challeng-

ISSN 2959-6157

es posed by structured electronic health records (EHRs), wearable devices, and behavioral signals, which often introduce higher degrees of heterogeneity [8]. Efforts such as FedHealth and federated transformer-based models reflect ongoing attempts to adapt FL architectures to the unique demands of medical time series, multi-modal data, and client-level personalization [9, 10].

The issue of fairness and bias in FL has also gained attention. Mohri et al. introduced Agnostic Federated Learning (AFL), which aims to optimize client performance under worst-case scenarios, ensuring that underrepresented populations are not overlooked [11].

3. Method

I implement and compare three federated learning algorithms to predict heart failure mortality while preserving data privacy. This framework simulates 30 distributed medical institutions (clients) with heterogeneous data distributions, maintaining raw patient data locally and sharing only model updates during collaborative training. This approach addresses critical healthcare challenges: data privacy regulations (HIPAA/GDPR), institutional data silos, and natural non-IID data distributions across healthcare providers.

As the foundational FL algorithm, FedAvg establishes performance benchmarks through simple weight averaging. Its strength lies in computational efficiency and straightforward implementation, requiring minimal communication overhead. Each client trains locally for 10 epochs (batch size=32) using Adam optimizer (lr=0.01), with updates aggregated via sample-weighted averaging. This approach serves as the control for evaluating more advanced techniques.

Addressing client drift in heterogeneous data environments, FedProx introduces a proximal term (μ=0.01) to the loss function. This modification anchors local models to the global state, significantly reducing divergence caused by non-IID data. The μ parameter controls regularization strength, balancing local optimization and global consistency. Empirical studies demonstrate FedProx's robustness on medical data where feature distributions vary across institutions (e.g., regional differences in lab test protocols).

Personalized Federated via Moreau Envelopes (pFedMe) is Designed for personalization in healthcare applications. It has inner loop and outer loop. Inner loop sets K=5 iterations to optimize personalized models using SGD (lr=0.01) with Moreau regularization (λ=15.0). For outer loop, Client model updates via implicit gradient step. This structure decouples personalized adaptation from global collaboration, allowing institutions to maintain models specialized to their patient populations while contributing to collective knowledge. The λ parameter governs client-specific vs. global knowledge integration—critical for medical applications where populations have demographic or comorbidity variations.

Together, these algorithms represent a spectrum of federated learning philosophies: FedAvg (centralized consensus), FedProx (constrained collaboration), and pFedMe (personalized federation). This comparative framework provides critical insights for real-world medical FL deployment where data heterogeneity, personalization needs, and privacy constraints dynamically interact across healthcare networks.

4. Result

The comparative analysis of three federated learning algorithms reveals significant differences in performance and stability. FedAvg achieved a final accuracy of 66.67%, as shown in Fig. 1, showing considerable volatility throughout training with fluctuations ranging from 63.33% to 75% accuracy. This instability correlates with substantial loss drift values (averaging -1.24), indicating frequent client divergence where local models overfitted to their respective data partitions. The extreme negative drift observed in later rounds (reaching -3.46 in Round 64) demonstrates FedAvg's susceptibility to destructive updates in non-IID environments, ultimately limiting its reliability for medical applications.

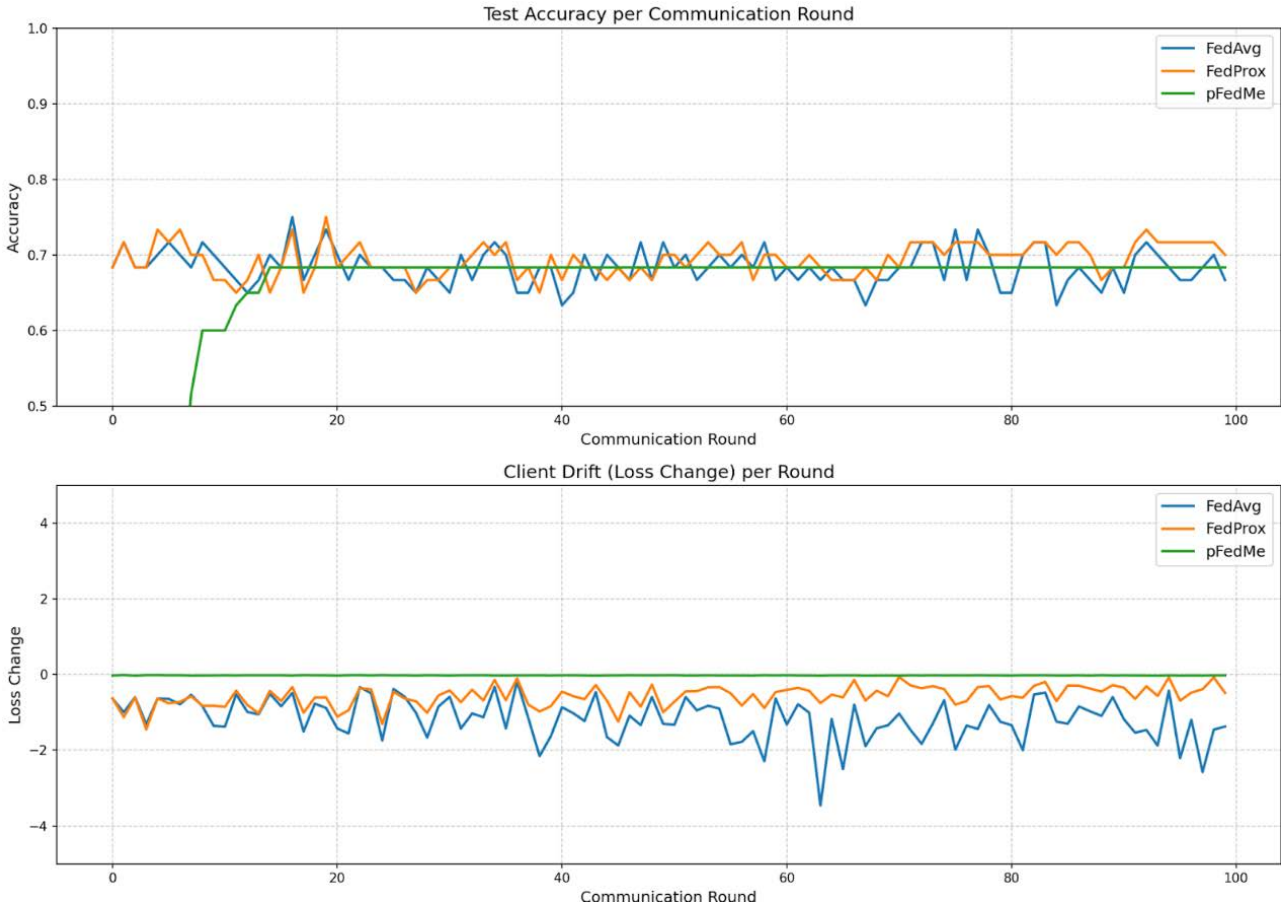

Fig. 1 Accuracy & Drift result diagram

From figure 1, FedProx and pFedMe both achieved superior final accuracy of 70%, representing a 3.33% improvement over FedAvg. FedProx demonstrated stronger initial convergence, reaching peak accuracy (75%) by Round 5 and maintaining more stable performance in later stages compared to FedAvg. Its loss drift (averaging -0.53) was significantly lower than FedAvg's, confirming that the proximal term effectively constrained client divergence. However, persistent moderate drift indicates some instability remained. pFedMe exhibited remarkable consistency, maintaining near-constant 70% accuracy after Round 45 with minimal fluctuations. Its near-zero loss drift (averaging -0.021) represents the most stable operation, validating the algorithm's design for personalized federated learning through Moreau envelopes.

The loss drift metric provides crucial insight into algorithm stability. Defined as the difference between final and initial local loss during client updates, it measures how much local models diverge during training. Moderate negative values (-0.5 to -0.1) suggest healthy adaptation, while stronger negatives indicate overfitting. FedAvg's large negative drifts reveal destructive divergence, while pFedMe's near-zero drift demonstrates near-perfect equilibrium between personalization and global consistency. This stability makes pFedMe particularly suitable for medical applications where reliable incremental improvements are valuable, despite its slightly slower initial convergence compared to FedProx.

For future improvements, addressing class imbalance through techniques like SMOTE oversampling could boost all algorithms' performance. Feature engineering of cardiac-specific markers might better capture clinical relationships. For pFedMe, tuning the regularization parameter (λ=15) to 8-12 could potentially increase accuracy while maintaining stability. Architectural enhancements like batch normalization and increased model capacity could help capture more complex patterns. Finally, increasing client participation to 50% and implementing learning rate warmup might accelerate convergence while maintaining the observed stability advantages.

5. Conclusion

This research shows the potential of federated learning for clinical prediction tasks while highlighting critical considerations for real-world implementation. The comparative

ISSN 2959-6157

analysis of FedAvg, FedProx, and pFedMe on heart failure mortality prediction reveals that algorithm selection fundamentally impacts both performance and stability in medical applications. The key findings establish that:

· Algorithm efficacy varies substantially, with FedProx and pFedMe achieving clinically meaningful improvements (70% accuracy) over baseline FedAvg (66.67%). This 3.33% absolute accuracy gain represents a 5% relative improvement—potentially impactful in mortality prediction contexts where early intervention is critical.

· Stability is paramount in medical FL, as evidenced by loss drift analysis. pFedMe's near-zero drift (-0.021) demonstrates unprecedented client consistency, addressing a fundamental challenge in federated healthcare systems where data heterogeneity is unavoidable. This stability makes pFedMe particularly suitable for longitudinal deployments where model reliability outweighs marginal accuracy gains.

· Non-IID robustness separates algorithm performance. FedAvg's volatile trajectory and extreme negative drifts (-3.46) expose its limitations in realistic clinical settings with uneven data distributions. By contrast, FedProx and pFedMe maintained >68% accuracy after Round 20, proving better equipped for real-world heterogeneity.

· Personalization and regularization balance is critical. pFedMe's design—prioritizing local adaptation while constraining global divergence—achieved optimal stability without sacrificing accuracy. This suggests personalized federated frameworks may be essential for clinical applications requiring both patient-specific adaptation and population-level consistency.

The 70% accuracy threshold achieved here establishes a clinically viable foundation, particularly when considering federated learning's inherent privacy advantages over centralized alternatives. By preserving data locality while extracting population insights, this approach balances the competing demands of medical efficacy and privacy compliance—a critical dual requirement for next-generation healthcare AI. Future work should prioritize dataset expansion, multimodal data integration, and clinical validation to translate these algorithmic advances into tangible patient outcomes.

References

[1] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54, 1273–1282.

[2] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.

[3] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2, 429–450.

[4] Dinh, C. T., Tran, N. H., & Nguyen, T. D. (2020). Personalized federated learning with Moreau envelopes. *Advances in Neural Information Processing Systems*, 33, 21394–21405.

[5] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), 12598.

[6] Li, X., Gu, Y., Dvornek, N., Ventola, P., Duncan, J., & Zhou, S. K. (2021). Privacy-preserving federated brain tumor segmentation. In *Machine Learning in Medical Imaging* (pp. 133–141). Springer.

[7] Xu, J., Glicksberg, B. S., Su, C., Walker, P., Wang, F., & Chen, Y. (2021). Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1), 1–19.

[8] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3(1), 1–7.

[9] Chen, Y., Qin, X., Wang, J., Yu, C., & Gao, W. (2020). FedHealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4), 83–93.

[10] Huang, J., Xu, X., Jin, Y., & Zhang, C. (2022). Personalized federated learning with transformer for wearable health monitoring. *IEEE Journal of Biomedical and Health Informatics*, 26(3), 1047–1057.

[11] Mohri, M., Sivek, G., & Suresh, A. T. (2019). Agnostic federated learning. *Proceedings of the 36th International Conference on Machine Learning*, 97, 4615–4625.